

FORM PTO-1390 (Modified)
(REV 11-2000)

U.S. DEPARTMENT OF COMMERCE PATENT AND TRADEMARK OFFICE

ATTORNEY'S DOCKET NUMBER

**TRANSMITTAL LETTER TO THE UNITED STATES
DESIGNATED/ELECTED OFFICE (DO/EO/US)
CONCERNING A FILING UNDER 35 U.S.C. 371**

A32550-PCT USA

U.S. APPLICATION NO. (IF KNOWN, SEE 37 CFR

10/018108

INTERNATIONAL APPLICATION NO.
PCT/US00/40238

INTERNATIONAL FILING DATE
19 June 2000

PRIORITY DATE CLAIMED
18 June 1999

TITLE OF INVENTION

SYSTEM AND METHOD FOR DETECTING TEXT SIMILARITY OVER SHORT PASSAGES

APPLICANT(S) FOR DO/EO/US

KLAVANS, Judith L.; ESKIN, Eleazar; and HATZIVASSILOGLU, Vasileios

Applicant herewith submits to the United States Designated/Elected Office (DO/EO/US) the following items and other information:

1. ☒ This is a **FIRST** submission of items concerning a filing under 35 U.S.C. 371.
2. ☐ This is a **SECOND** or **SUBSEQUENT** submission of items concerning a filing under 35 U.S.C. 371.
3. ☐ This is an express request to begin national examination procedures (35 U.S.C. 371(f)). The submission must include items (5), (6), (9) and (24) indicated below.
4. ☒ The US has been elected by the expiration of 19 months from the priority date (Article 31).
5. ☒ A copy of the International Application as filed (35 U.S.C. 371 (c) (2))
 - a. ☒ is attached hereto (required only if not communicated by the International Bureau).
 - b. ☐ has been communicated by the International Bureau.
 - c. ☐ is not required, as the application was filed in the United States Receiving Office (RO/US).
6. ☐ An English language translation of the International Application as filed (35 U.S.C. 371(c)(2)).
 - a. ☐ is attached hereto.
 - b. ☐ has been previously submitted under 35 U.S.C. 154(d)(4).
7. ☒ Amendments to the claims of the International Application under PCT Article 19 (35 U.S.C. 371 (c)(3))
 - a. ☐ are attached hereto (required only if not communicated by the International Bureau).
 - b. ☐ have been communicated by the International Bureau.
 - c. ☐ have not been made; however, the time limit for making such amendments has NOT expired.
 - d. ☒ have not been made and will not be made.
8. ☐ An English language translation of the amendments to the claims under PCT Article 19 (35 U.S.C. 371(c)(3)).
9. ☐ An oath or declaration of the inventor(s) (35 U.S.C. 371 (c)(4)).
10. ☐ An English language translation of the annexes of the International Preliminary Examination Report under PCT Article 36 (35 U.S.C. 371 (c)(5)).
11. ☐ A copy of the International Preliminary Examination Report (PCT/IPEA/409).
12. ☒ A copy of the International Search Report (PCT/ISA/210).

Items 13 to 20 below concern document(s) or information included:

13. ☐ An Information Disclosure Statement under 37 CFR 1.97 and 1.98.
14. ☐ An assignment document for recording. A separate cover sheet in compliance with 37 CFR 3.28 and 3.31 is included.
15. ☐ A **FIRST** preliminary amendment.
16. ☐ A **SECOND** or **SUBSEQUENT** preliminary amendment.
17. ☐ A substitute specification.
18. ☐ A change of power of attorney and/or address letter.
19. ☐ A computer-readable form of the sequence listing in accordance with PCT Rule 13ter.2 and 35 U.S.C. 1.821 - 1.825.
20. ☐ A second copy of the published international application under 35 U.S.C. 154(d)(4).
21. ☐ A second copy of the English language translation of the international application under 35 U.S.C. 154(d)(4).
22. ☒ Certificate of Mailing by Express Mail
23. ☒ Other items or information:

Forms PCT/IPEA/401/, Forms PCT/ISA/210/220; Forms PCT/IB/301/304/308, a postcard and a check in the amount of \$355.

Express Mail No.: EF377397927US

Date of Deposit: December 13, 2001

U.S. APPLICATION NO. (IF KNOWN, SEE 37 CFR 1.101) 10/018108		INTERNATIONAL APPLICATION NO. PCT/US00/40238		ATTORNEY'S DOCKET NUMBER A32550-PCT USA	
---	--	--	--	---	--

24. The following fees are submitted:.				CALCULATIONS PTO USE ONLY	
BASIC NATIONAL FEE (37 CFR 1.492 (a) (1) - (5)) : <input type="checkbox"/> Neither international preliminary examination fee (37 CFR 1.482) nor international search fee (37 CFR 1.445(a)(2)) paid to USPTO and International Search Report not prepared by the EPO or JPO \$1000.00 <input checked="" type="checkbox"/> International preliminary examination fee (37 CFR 1.482) not paid to USPTO but International Search Report prepared by the EPO or JPO \$860.00 <input type="checkbox"/> International preliminary examination fee (37 CFR 1.482) not paid to USPTO but international search fee (37 CFR 1.445(a)(2)) paid to USPTO \$710.00 <input type="checkbox"/> International preliminary examination fee (37 CFR 1.482) paid to USPTO but all claims did not satisfy provisions of PCT Article 33(1)-(4) \$690.00 <input type="checkbox"/> International preliminary examination fee (37 CFR 1.482) paid to USPTO and all claims satisfied provisions of PCT Article 33(1)-(4) \$100.00 <div style="text-align: right;">ENTER APPROPRIATE BASIC FEE AMOUNT =</div>				\$710.00	
Surcharge of \$130.00 for furnishing the oath or declaration later than months from the earliest claimed priority date (37 CFR 1.492 (e)). <input type="checkbox"/> 20 <input type="checkbox"/> 30				\$0.00	
CLAIMS	NUMBER FILED	NUMBER EXTRA	RATE		
Total claims	16 - 20 =	0	x \$18.00	\$0.00	
Independent claims	2 - 3 =	0	x \$78.00	\$0.00	
Multiple Dependent Claims (check if applicable). <input type="checkbox"/>				\$0.00	
TOTAL OF ABOVE CALCULATIONS =				\$710.00	
<input checked="" type="checkbox"/> Applicant claims small entity status. (See 37 CFR 1.27). The fees indicated above are reduced by 1/2.				\$355.00	
SUBTOTAL =				\$355.00	
Processing fee of \$130.00 for furnishing the English translation later than months from the earliest claimed priority date (37 CFR 1.492 (f)). <input type="checkbox"/> 20 <input type="checkbox"/> 30				\$0.00	
TOTAL NATIONAL FEE =				\$355.00	
Fee for recording the enclosed assignment (37 CFR 1.21(h)). The assignment must be accompanied by an appropriate cover sheet (37 CFR 3.28, 3.31) (check if applicable). <input type="checkbox"/>				\$0.00	
TOTAL FEES ENCLOSED =				\$355.00	
				Amount to be: refunded	\$
				charged	\$

a. ☒ A check in the amount of **\$355.00** to cover the above fees is enclosed.

b. ☐ Please charge my Deposit Account No. _____ in the amount of _____ to cover the above fees
A duplicate copy of this sheet is enclosed.

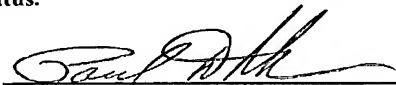
c. ☒ The Commissioner is hereby authorized to charge any additional fees which may be required, or credit any overpayment to Deposit Account No. **02-4377** A duplicate copy of this sheet is enclosed.

d. ☐ Fees are to be charged to a credit card. **WARNING:** Information on this form may become public. **Credit card information should not be included on this form.** Provide credit card information and authorization on PTO-2038.

NOTE: Where an appropriate time limit under 37 CFR 1.494 or 1.495 has not been met, a petition to revive (37 CFR 1.137(a) or (b)) must be filed and granted to restore the application to pending status.

SEND ALL CORRESPONDENCE TO:

Henry Tang
Baker Botts LLP
30 Rockefeller Plaza
New York, NY 10112-0228


 SIGNATURE

Paul D. Ackerman
 NAME

39,891
 REGISTRATION NUMBER

December 13, 2001
 DATE

SYSTEM AND METHOD FOR DETECTING TEXT SIMILARITY OVER SHORT PASSAGES

FIELD OF THE INVENTION

The present invention relates generally to natural language processing and
5 more particularly relates to a system and method for determining the similarity of text
in short passages.

BACKGROUND OF THE INVENTION

With the growing volume of textual information, such as newspaper articles,
magazines, Internet articles, and the like, there is a growing need to automatically
10 cluster and/or classify such documents and determine whether groups of documents
express similarities or not. For the most part, research in this area has focused on
detecting similarity between documents and large segments of text or between a short
query phrase and one or more documents.

While effective techniques have been developed for document clustering and
15 classification which depend on inter-document similarity measures, these techniques
generally rely only on shared words, or occasionally on collocation of words. Such
techniques are applicable when large units of text, such as full documents, are
compared. In this case, there is generally sufficient overlap to detect similarity in the
documents and/or document segments. However, when the units of text are small, for
20 example a paragraph or abstract, such simple surface matching of words and phrases
is far more prone to error. In the case of small text units, the sample size is reduced
and the number of potential matches is reduced accordingly. Thus, there remains a
need for improved techniques for detecting similarities between small text units.

A further problem with known techniques for detecting similarity is that the
25 conventional notions of similarity which are applicable to large text samples, such as
documents and large text segments, do not provide sufficient measures of similarity
for measuring similarity in small text segments. Standard notions of similarity
generally involve the creation of a vector or profile of characteristics of a text

fragment and determine a conceptual distance between vectors on the basis of frequencies. Features typically include stemmed words, although multi-word units and collocations also have been used. Typological characteristics, such as thesaural features, have also been used to calculate features. The difference between vectors for one text unit (usually a query) and another text unit (usually a document) then determines closeness or similarity of the text units.

In some cases, the text units are represented as vectors of sparse n-grams of word occurrences and learning is applied over those vectors. Though effective in the context of large document comparisons, a more fine-grained distinction for similarity measures is required to properly characterize the similarity of two small text segments.

SUMMARY OF THE INVENTION

It is an object of the present invention to provide systems and methods for detecting similarity between two or more small text segments.

A method for determining similarity in short text segments in accordance with the present invention includes the steps of determining common primitive features in the text segments, determining common composite features in the text segments and then calculating a similarity measure based upon the primitive and composite features. The primitive features can be selected from the group including common single words, common noun phrases, synonyms, common semantic classes of verbs, and common proper nouns. The composite features, which represent relationships between and among the primitive features, can be selected from the group including primitive feature order restrictions, primitive feature distance restrictions, and primitive type restrictions.

Preferably, the step of determining common primitive features can include the further steps of identifying common primitive features, assigning a value to the primitive features, and normalizing the feature values. Normalizing the values can include normalizing for text segment length and normalizing for the frequency of primitive feature occurrence. Similarly, determining composite features generally includes identifying the composite features, assigning a value to the composite

features, and normalizing the feature values. Again, normalization of the feature values can include normalizing for text segment length and normalizing for the frequency of feature occurrence.

BRIEF DESCRIPTION OF THE DRAWING

5 Further objects, features and advantages of the invention will become apparent from the following detailed description taken in conjunction with the accompanying figures showing illustrative embodiments of the invention, in which

Figure 1 is a flow chart illustrating an overview of a present method for comparing small text segments;

10 Figure 2 is a flow chart illustrating the step of defining similarity for small text segments in accordance with the present methods;

Figure 3 is a flow chart illustrating the process of computing primitive features for use in detecting similarity in small text segments;

15 Figure 4 is a flow chart illustrating the process of calculating composite features for use in detecting similarity of small text segments in accordance with the present methods;

Figure 5 is a block diagram of a software system topology for determining similarity in small text segments in accordance with the present methods;

Figure 6 is an illustration of exemplary short text segments;

20 Figure 7 is a diagram illustrating a composite feature match between two of the short text segments provided in Figure 6 using a "same order" rule;

Figure 8 is a diagram illustrating a composite feature match between two of the short text segments provided in Figure 6 using a "within distance" rule; and

25 Figure 9 is a diagram illustrating a composite feature match between two of the short text segments provided in Figure 6 using a "primitive type" rule.

Throughout the figures, the same reference numerals and characters, unless otherwise stated, are used to denote like features, elements, components or portions of the illustrated embodiments. Moreover, while the subject invention will now be described in detail with reference to the figures, it is done so in connection with the
30 illustrative embodiments. It is intended that changes and modifications can be made

to the described embodiments without departing from the true scope and spirit of the subject invention as defined by the appended claims.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

Figure 1 is a flow chart illustrating an overview of the process used in the present invention for detecting similarity in small text segments. As previously noted, a problem in the prior art is that the definition of similarity commonly used for large text segments, such as documents, is not sufficiently refined to provide an adequate measure of similarity when comparing small text segments. Generally, small text segments refer to sentences, phrases and short paragraphs.

Referring to Figure 1, in step 100 a definition of similarity for small text segments is provided. From this definition, the method proceeds to identify primitive features of the small text segments and determine feature values for the primitive features (step 105). Primitive features are those which generally compare simple parts of speech and text, such as single words, word categories, or phrases such as noun phrases, synonyms, verb class and proper nouns. In addition to primitive features, the process can identify composite features of the short-text segments and determine composite feature values (step 110). Composite features are those which compare relationships among two or more primitive features. Once primitive features and composite features have been identified and given an appropriate value, a machine learning algorithm is applied to classify small text segments as similar or not similar (step 115).

Figure 2 is a flow chart which illustrates the process of establishing an appropriate definition of similarity for small text segments. In general, two text units can be considered as similar if they share the same focus on a common concept, actor, object or action. In addition, the common actor or object definition must perform or be subjected to the same action or be the subject of the same description. This is exemplified in the flow chart of Figure 2, where two small text segments are selected from a body of text and are analyzed. If the two text segments relate to a common concept (step 205), then further analysis is performed to see if the common concept relates to the same action (step 210) or relates to the same description (step 215).

Similar tests are performed to determine if the two text segments relate to a common actor (step 220) or to a common object (step 225). If there is no common concept, actor or object, the text segments are considered not similar (step 235). Similarly, for those text segments which do refer or relate to a common concept, actor or object, those segments will still be found not similar unless they also relate to a common action or involve the same description. Thus, for short text segments to be similar, they must contain a common concept, actor, or object which is also the subject of a common action or description. The comparisons in steps 205, 220 and 225 can be the basis for primitive features 240. Those relationships between primitive features which are identified in steps 210, 215 can be referred to as composite features 245.

While Figure 2 is illustrated as a sequential process, it represents a decision tree involved in a definition of similarity of two short text segments as applied in the present invention which can also be performed in a largely parallel manner. For example, decisions 205, 220 and 225 can be performed concurrently as can decisions 210 and 215. Using this definition of similarity for small text segments, a feature-based process can be employed which compares primitive and composite features of short text segments to determine if the definition is satisfied for two or more given input text segments.

Figure 3 is a flow chart which illustrates a method for extracting and scaling primitive features in accordance with the present invention. The text segments are compared for a level of commonality, including determining whether there is a common single word (step 305), a common noun phrase (step 310), whether two words in the phrases are synonyms (step 315), whether the phrases include verbs having a common semantic class (step 320), and whether a common proper noun can be found in the two phrases (step 325). If none of these conditions are satisfied for the applied small text segments, there is no primitive feature common to these two text segments (step 327). When a primitive feature has been identified, e.g., one of the conditions in steps 305 through 325 is satisfied, a feature value is assigned to that primitive feature. Preferably, the values which are assigned to the features are determined by a machine learning algorithm, such as RIPPER, which is trained using a suitable training corpus. RIPPER is a widely-used and effective rule induction

system which is available from AT&T Laboratories and is described by Cohen in "Learning Trees and Rules with Set-Valued Features, Proceedings of the Fourteenth National Conference on Artificial Intelligence, American Association on Artificial Intelligence, 1996, which is incorporated by reference. It has been found that a subset of a corpus of 264 paragraphs which have been manually tagged by human readers as similar or not similar can be used to establish a feature rule set for RIPPER which is then suitable for assigning values to the features identified in the text segments. The particular training corpus and learned rule set will generally vary depending on the desired application. The values assigned will vary based on properties of the machine learning algorithm and training corpus. After feature values are assigned in step 330, these values can be normalized based on text length (step 335) and/or noted frequency of occurrence (step 340). Though normalization is optional, it is a desirable step to provide uniform and accurate results across varying types of text and length of text segments.

Primitive features provide a baseline indication of similarity. To further refine the notion of similarity in small text segments, relationships among primitive features, referred to as composite features, can also be evaluated. Referring to Figure 4, a method of evaluating composite features is illustrated. Composite features are those features which identify relationships among primitive feature pairs. Generally, composite features are defined by placing different forms of restrictions on participating primitive feature pairs. Referring to Figure 4, the primitive features identified in each of the small text segments are applied to a test layer 400 where various feature relationships are evaluated. The relationships illustrated in test layer 400 are exemplary in nature and are not intended to illustrate an exhaustive list of possible relationships. It will be appreciated that an large number of relationships between and among primitive features can be used to establish composite features.

For example, one type of feature relationship for composite features can be that the primitives occur in the same order in each of the text samples (step 405). This is illustrated by example in Figure 7. Figure 6 provides three short text segments to be compared. Figure 7 illustrates a match according to the "same order" composite feature rule. In Figures 7-9, primitive features are identified by shading and the

relationships which form the composite features are illustrated by connecting lines. In the case illustrated in Figure 7 the primitive features {two, contact} appear in the same order in text segments Figure 6 (a) and 6 (b) from Figure 6.

Another possible relationship is that two pairs of primitive elements are
5 required to occur within a certain distance in both text segments. The maximum distance between the primitive elements which would satisfy the relationship can be a variable or a predetermined constant (step 410). Referring to Figure 8, an example of a positive match for the "within distance" composite feature rule is provided, given that the distance, n , is set to a value less than three. In Figure 8, although the primitive
10 features {contact, lost} do not appear in the same order, they occur within n words of each other ($n < 3$ in this case).

Yet another exemplary relationship can be that the two text segments include the same primitive feature types. For example, one primitive feature can be restricted to a simplex noun phrase while the other to a verb. In such a case, two noun phrases,
15 one from each text unit, must match according to the rule for matching simplex noun phrases and two verbs must match according to the applied rules of verb primitives (e.g., sharing the same semantic class). This is illustrated in Figure 9 where the primitive feature "An OH-58 helicopter" is deemed a simplex noun phrase match with
"the helicopter" and both phrases include a common verb, "lost".

20 By matching primitive feature types, a simple grammatical relationship is determined in the text segments. Returning to Figure 4, for each condition that is satisfied in test layer 400, feature values are assigned to those composite features identified (step 420). The feature values are assigned by a machine learning algorithm, such as RIPPER, which has been trained on a suitable training corpus. As
25 in the case of primitive features, optionally, the feature values assigned to the composite feature can be normalized for text length and relative occurrence of the primitive feature or composite feature (steps 425, 430, respectively). Once both primitive features and composite features of the small text segments have been identified, a machine learning algorithm is applied to determine a similarity value
30 between the text segments (step 435). The machine learning algorithm can perform a rule-based analysis to determine similarity. Alternatively, a simpler algorithm can be

used to determine similarity by comparing the total feature value of the text segments being compared to a predetermined threshold value.

Figure 5 is a block diagram of an exemplary software system for conducting the method described in connection with Figures 1-4. The system is generally
 5 implemented in software for a general purpose computer, such as a personal computer or work station. The system includes a main processing section 500. One or more interface modules 510 are included for receiving text input for the text segments to be compared and for providing the text segments to the main processing section 500. The text input can be provided by a number of sources, including but not limited to,
 10 computer readable memory, hard disks, optical disks, network databases, on-line sources, manual keyed input and the like. Based on the desired text source and input mechanism, one skilled in the art can provide appropriate text input interface module 510 hardware and software.

The main processing section 500 is also operatively coupled to a training
 15 corpus 515, which is generally stored in computer readable storage media. The main processing section 500 is generally programmed in a structured manner which calls various subprograms, library routines, and the like to perform the various functions described in accordance with Figures 1-4. The main processing section 500 can invoke the various subroutines sequentially (serial) or in a parallel, or batched,
 20 processing mode. The received text is generally passed to a preprocessing routine 520. The preprocessing routine cleans up the received text, such as by removing control characters from the text. The preprocessing routine also performs part-of-speech (POS) tagging, using known techniques, such as are available in the ALEMBIC tool set, described by Aberdeen et al. in "MITRE: Description of the
 25 Alembic System as used for MUC-6," Proceedings of the Sixth Message Understanding Conference, 1995, which is hereby incorporated by reference. ALEMBIC provides a set of data and language processing tools which identify the various parts of speech present in the small text segments.

Following text preprocessing, control is returned to the main processing
 30 section 500 which then preferably invokes a noun phrase comparison subroutine 525, such as LinkIt, to perform noun phrase comparison of step 310. LinkIt can be

employed to determine whether a common noun phrase is present in the applied text segments and for identifying simplex noun phrases and matching those that share the same noun head. The LinkIt tool is described by N. Wacholder in "Simplex NPs Clustered by Head: A Method for Identifying Significant Topics in a Document",
5 Proceedings of the Workshop on the Computational Treatment of Nominals, October 1998, which is hereby incorporated by reference in its entirety.

To determine if two segments include common proper nouns as required in step 325, the noun comparison algorithm can also be used to match those nouns identified using the ALEMBIC toolset using various predetermined matching criteria.
10 Variations on proper noun matching can include restricting the proper noun type to a person, place or organization. Such subcategories can also be extracted using ALEMBIC's named entity finder.

Following noun phrase identification and matching, other routines for detecting primitive features can be employed. For example, to perform step 305 and
15 determine whether common single word primitive features exist between two text segments, a word co-occurrence detection sub-routine 540 can be called by the main program 500. Variations of the word co-occurrence operation can restrict matching to cases where the parts of speech of the words also match, or relax the comparison to cases where only the word stems of the two words are identical.

20 Similarly, to determine if two text segments include words which are synonyms, a synonym detection algorithm 530 can be called by the main processing routine 500. In this regard, a lexical database such as WordNet®, as described by G. Miller in "WordNet, An On-Line Lexical Database," International Journal of Lexicography, Vol. 3, No. 4 (1990), can be employed. WordNet provides sense
25 information and places words in sets of synonyms (synsets). Words that appear in the same synset are generally considered matches. Variations on this feature can be used to restrict the words being compared to a specific part-of-speech class.

To determine if two verbs present in the short text segments are of the same semantic class as set forth in step 320, a verb classifier and comparator algorithm 535
30 can be operatively coupled to the main processing section 500 and called by the main program. Semantic classes for verbs have been found to be useful for determining

document types and text similarity. This is discussed, for example, in “The Role of Verbs in Document Analysis” by J. Klavans et al., Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics, 1998, which is hereby incorporated by
 5 reference in its entirety. For those verbs which are found to have a common semantic class, e.g., communication, motion, agreement, argument, etc., those verbs are considered to match.

The program operating in main processing section 500 can also provide algorithms to normalize feature values for text lengths and relative occurrence of the
 10 primitive. To normalize feature values for text length, as set forth in step 335, each feature value can be normalized by the size of the textual segments in the pair. For example, for a pair of textual segments A and B, the feature values assigned are divided by a normalization value, N:

$$N = \sqrt{\text{Length}(A) \times \text{Length}(B)} \quad (1)$$

This operation removes any potential bias in favor of longer text segments. It is noted
 15 that the units involved in the lengths of A and the lengths of B are generally measured by a word count.

Normalization of feature values can also be based on the relative frequency of occurrence of each primitive feature. Such normalization is motivated by the general observation that infrequently matching primitive elements are likely to have a higher
 20 impact on similarity than primitives which match more frequently. Such normalization is similar to the document frequency component of the commonly employed TF*IDF calculation. In this case, each primitive feature is associated with a value which is equal to the number of textual units in which the primitive appeared in the corpus. For a primitive element which compares single words, this is the number
 25 of text segments which contain that word in the corpus; for a noun phrase, this is the number of textual units that contain noun phrases that share the same head; and similarly for other primitive types. We multiply each feature's value by:

$$\text{Log}\left(\frac{T}{N}\right) \quad (2)$$

where T is a number of textual segments and N is the number of textual segments containing the primitive. It is noted that since normalization for text length and frequency of occurrence are both optional operations, when these two normalization techniques are selectively applied, there are up to four variations of normalizations for each primitive feature. Of course, other normalization techniques may be added to, or substituted for, the two methods discussed herein.

The program in main processing section 500 generally employs a machine learning algorithm 545 to determine whether the text units match overall. A suitable machine learning algorithm is RIPPER, as disclosed by Cohen in "Learning Trees and Rules with Set-Valued Features, Proceedings of the Fourteenth National Conference on Artificial Intelligence, American Association on Artificial Intelligence, 1996, which is incorporated by reference. RIPPER is a widely-used and effective rule induction system. This RIPPER algorithm is trained over a corpus of manually marked pairs of text units continued in the training corpus 515. A suitable corpus was constructed using a subset of the Topic Detection and Tracking (TDT) corpus developed by NIST and DARPA. The TDT corpus is a collection of over 16,000 news articles from Reuters and CNN where many of the articles have been manually grouped into 25 categories each of which correspond to a single event. The selected corpus was formed using the Reuters' articles in five of the twenty five categories from randomly selected days. The resulting training corpus 515 contained 30 related articles. The 30 articles provided 264 paragraphs which were selected as the small text segments and resulted in 10,345 comparisons between segments.

Although use of a machine learning algorithm is preferred, other algorithms can also be used. For example, an algorithm can add the total value of composite features found in the text segments and compare this value against a similarity threshold. Similarly, although it is preferred to determine feature values based on the use of a machine learning algorithm, feature values can be predetermined based on human experience through the use of a look-up table. Alternatively, all features can be given a binary value and the similarity comparison can be determined based on a simple accumulated count of detected primary and composite features.

The present methods, while evaluated on a corpus of English language documents, are not language specific and are generally applicable to any language. Of course, the individual subroutines may require some alteration to accommodate the varied constructions found in different languages.

5 The methods for determining similarity in small text segments described herein form an important component in larger systems, such as document archiving systems and multi-document summarization systems.

10 Although the present invention has been described in connection with specific exemplary embodiments, it should be understood that various changes, substitutions and alterations can be made to the disclosed embodiments without departing from the spirit and scope of the invention as set forth in the appended claims.

CLAIMS

1. A method for determining similarity in short text segments comprising:
determining common primitive features in the text segments;
determining common composite features in the text segments;
5 and
calculating a similarity measure based upon said primitive and
composite features.
2. The method for determining similarity as defined by claim 1, wherein
10 said primitive features are selected from the group including common single word,
common noun phrase, synonyms, common semantic class of verbs, and common
proper nouns.
3. The method for determining similarity as defined by claim 1, wherein
15 said composite features are selected from the group including primitive feature order
restrictions, primitive distance restrictions, and primitive type restrictions.
4. The method for determining similarity as defined by claim 1, wherein
said step of determining common primitive features includes:
identifying common primitive features;
assigning a value to said primitive features; and
20 normalizing said value.
5. The method for determining similarity as defined by claim 4, wherein
said step of normalizing includes at least one of normalizing for text segment length
and normalizing for frequency of primitive occurrence.
6. The method for determining similarity as defined by claim 1, wherein
25 said step of determining common composite features includes:
identifying common primitive features;

assigning a value to said primitive features; and
normalizing said value.

7. The method for determining similarity as defined by claim 6, wherein
said step of normalizing includes at least one of normalizing for text segment length
5 and normalizing for frequency of primitive occurrence.

8. A system for determining similarity in short text segments comprising:
an interface circuit for receiving text segments for comparison;
a main processing section, the main processing section being
operatively couple to the interface circuit and operating under the control of a
10 computer program, the program performing operations to determine common
primitive features in the text segments, determine common composite features in the
text segments; calculate a similarity measure based upon said primitive and
composite features, and provide an output indicative of the similarity measure.

15 9. The system for determining similarity as defined by claim 8, wherein
said primitive features are selected from the group including common single word,
common noun phrase, synonyms, common semantic class of verbs, and common
proper nouns.

20 10. The system for determining similarity as defined by claim 8, wherein
said composite features are selected from the group including primitive feature order
restrictions, primitive distance restrictions, and primitive type restrictions.

11. The system for determining similarity as defined by claim 8, wherein
the processing operation of determining common primitive features includes:
identifying common primitive features;
25 assigning a value to said primitive features; and
normalizing said value.

12. The system for determining similarity as defined by claim 11, wherein the processing operation of normalizing includes at least one of normalizing for text segment length and normalizing for frequency of primitive occurrence.

13. The system for determining similarity as defined by claim 8, wherein
5 said processing operation for determining common composite features includes:
identifying common primitive features;
assigning a value to said primitive features; and
normalizing said value.

14. The system for determining similarity as defined by claim 13, wherein
10 said processing operation for normalizing includes at least one of normalizing for text segment length and normalizing for frequency of primitive occurrence.

15. The system for determining similarity as defined by claim 8, wherein the computer program includes a noun phrase identification subroutine, a synonym detection subroutine, a verb classifier subroutine and a word co-occurrence
15 subroutine.

16. The system for determining similarity as defined by claim 8, further comprising a computer readable training corpus, and wherein the computer program includes a machine learning algorithm operatively coupled to the training corpus for learning and applying a rule set for determining similarity in small text segments.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
28 December 2000 (28.12.2000)

PCT

(10) International Publication Number
WO 00/79426 A1

(51) International Patent Classification⁷: G06F 17/21

(21) International Application Number: PCT/US00/40238

(22) International Filing Date: 19 June 2000 (19.06.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/139,930 18 June 1999 (18.06.1999) US

(71) Applicant (for all designated States except US): THE TRUSTEES OF COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK [US/US]; 116th Street and Broadway, New York, NY 10027 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): KLAVANS, Judith, L. [US/US]; 40 South Drive, Hastings-on Hudson, NY

10706 (US). ESKIN, Eleazar [—/US]; Columbia University, Shapiro Room 722, 116th Street and Broadway, New York, NY 10027 (US). HATZIVASSILOGLOU, Vasileios [—/US]; Columbia University, Shapiro Room 724, 116th Street and Broadway, New York, NY 10027 (US).

(74) Agents: TANG, Henry et al.; Baker Botts LLP, 30 Rockefeller Plaza, New York, NY 10112-0228 (US).

(81) Designated States (national): JP, US.

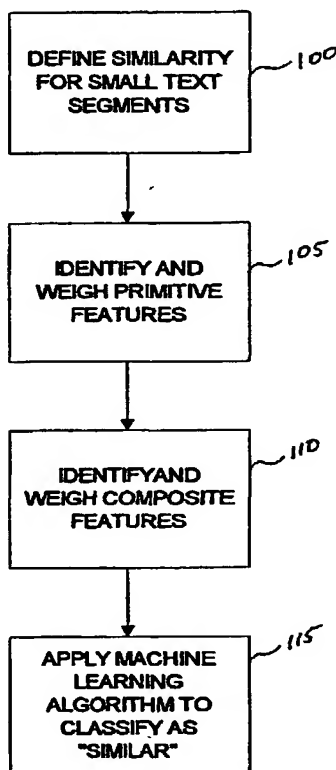
(84) Designated States (regional): European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published:

— With international search report.

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: SYSTEM AND METHOD FOR DETECTING TEXT SIMILARITY OVER SHORT PASSAGES



(57) Abstract: A system and method are provided for determining similarity in short text segments. The method provides a definition of similarity which is appropriate for the small text setting (100). Small text segments are compared to determine if there exist common primitive features, such as words, noun phrases, synonyms, verbs with a common semantic class, proper nouns and the like (105). From the primitive features identified, the small text segments are evaluated to determine whether composite features are present (110). Composite features are defined as predetermined relationships between primitive features. The common primitive features and composite features are applied as inputs to an appropriate machine learning algorithm which is trained to ascertain a similarity measure based on the primitive and composite features common to the text segments (115).

WO 00/79426 A1

FIG. 1

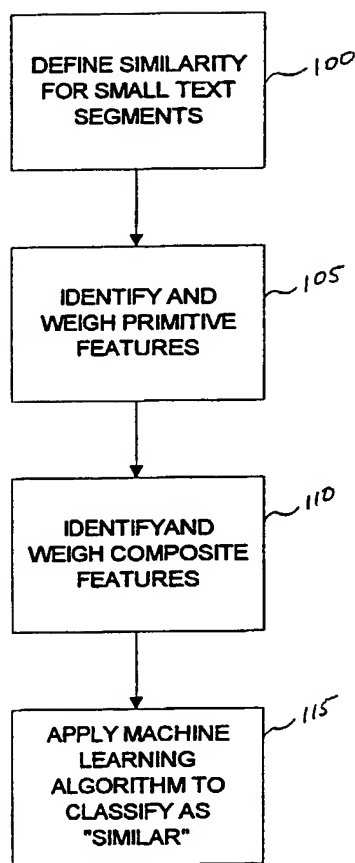


FIG. 2

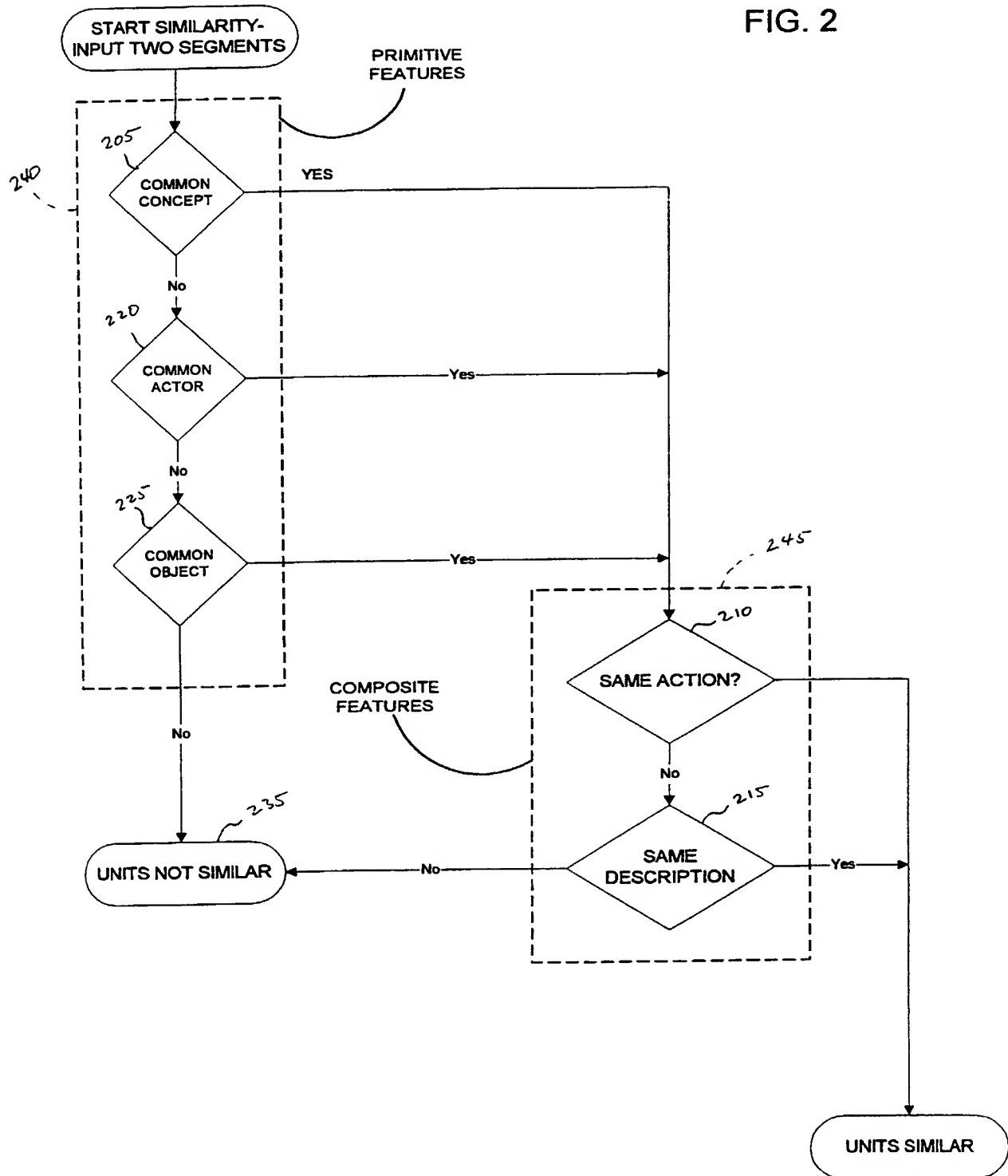
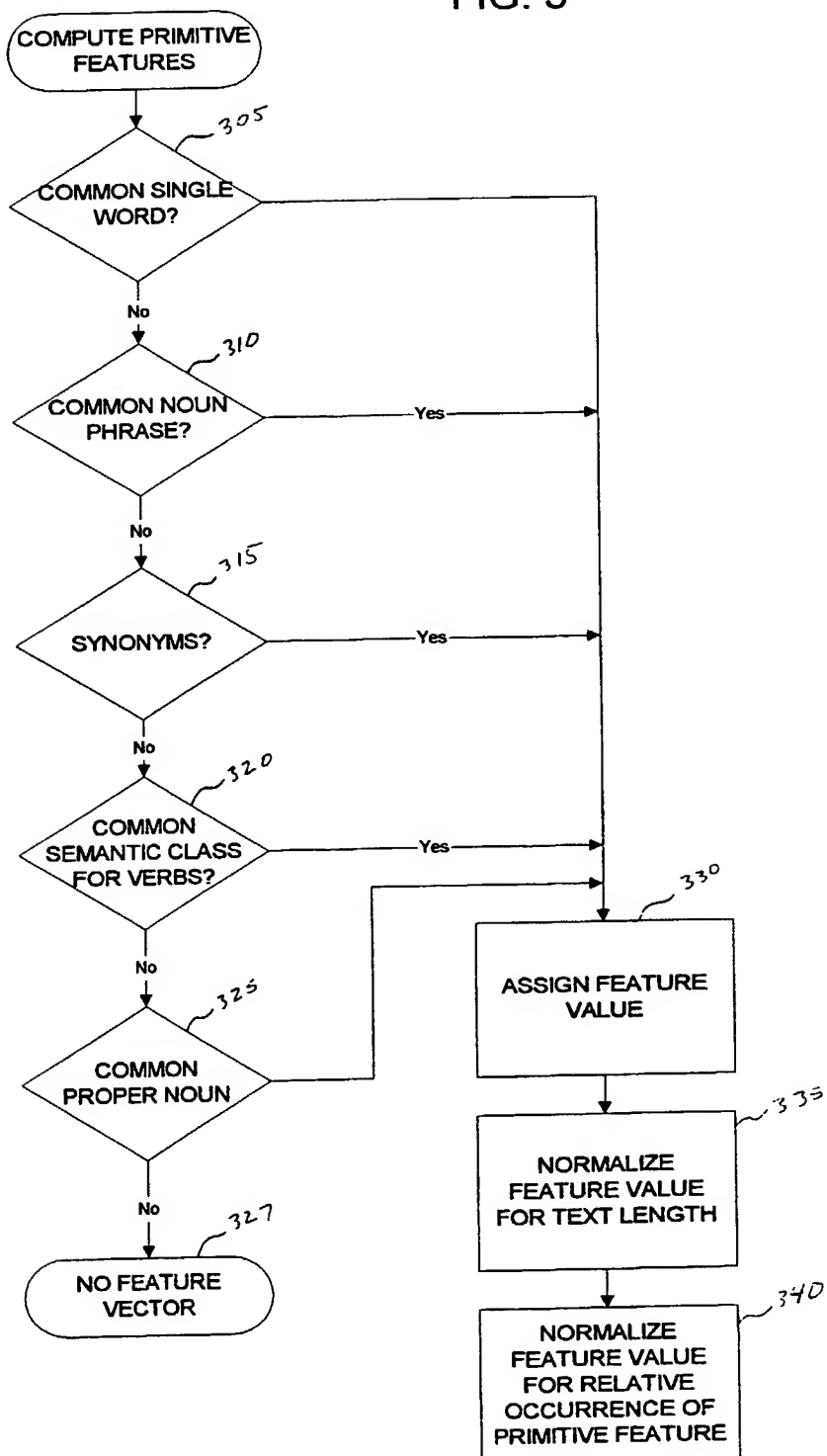


FIG. 3



10/018108

FIG. 4

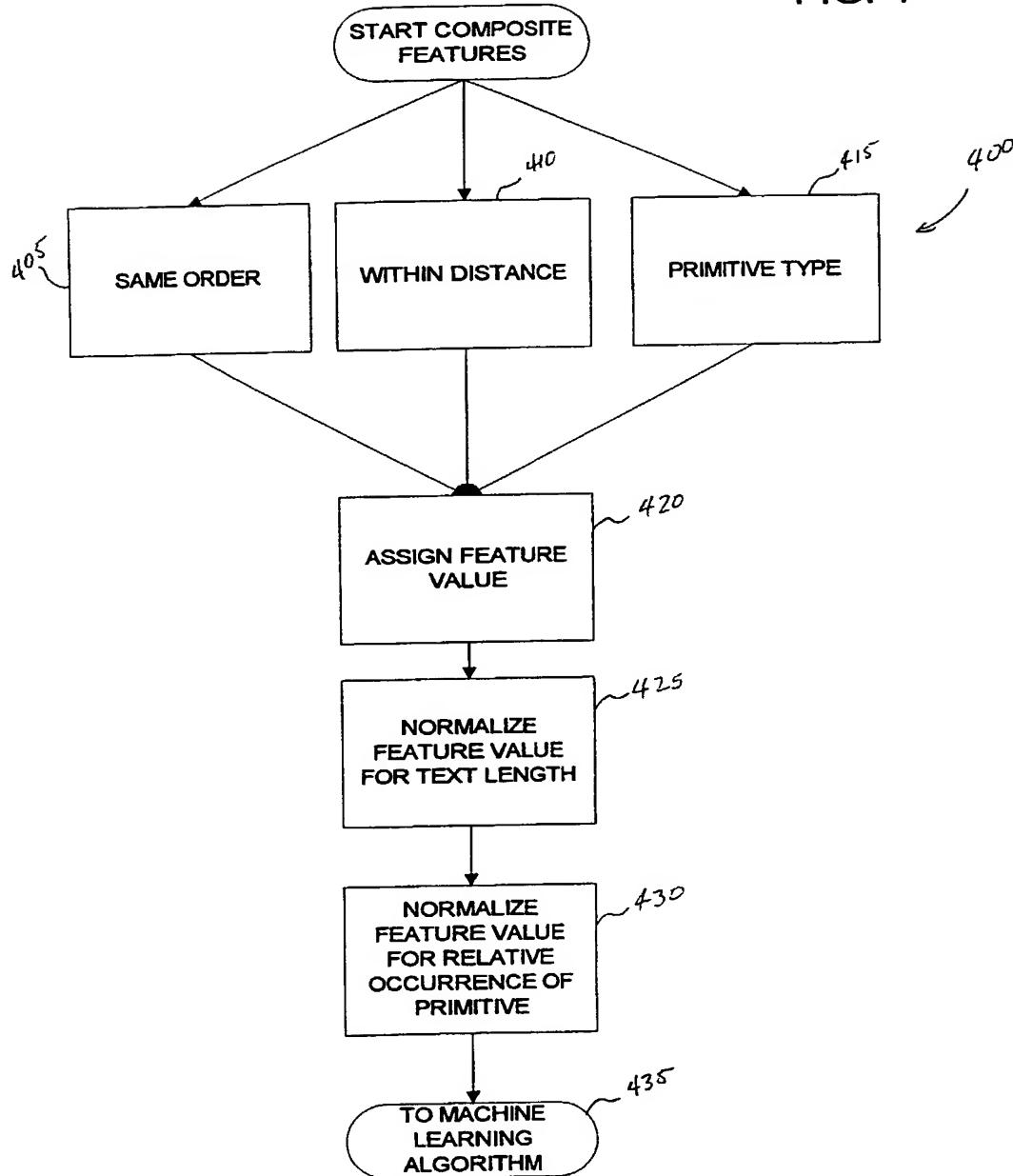
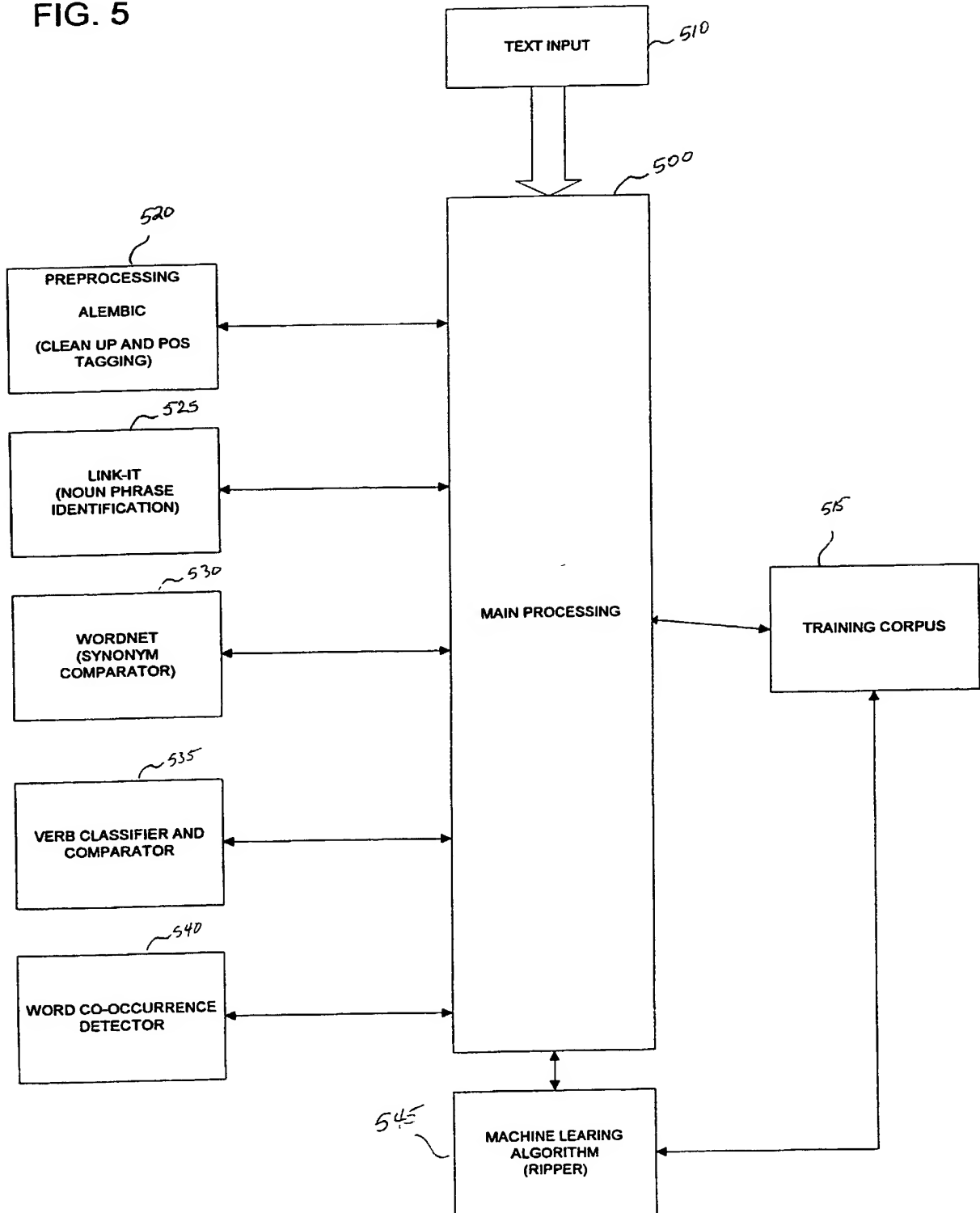


FIG. 5



10/018108

- Fig 6 (a) An OH-58 helicopter, carrying a crew of two, was on a routine training orientation when contact was lost at about 11:30 a.m. Saturday (9:30 p.m. EST Friday).
- Fig 6 (b) "There were two people on board," said Bacon. "We lost radar contact with the helicopter about 9:15 EST (0215 GMT)."
- Fig 6 (c) An OH-58 U.S. military scout helicopter made an emergency landing in North Korea at about 9:15 p.m. EST Friday (0215 GMT Saturday), the Defense Department said.

Figure 1: Input text units (from the TDT pilot—corpus, topic 11).

Fig 7

- (a) An OH-58 helicopter, carrying a crew of ~~two~~, was on a routine training orientation when ~~contact~~ was lost at about 11:30 a.m. Saturday (9:30 p.m. EST Friday).
- (b) "There were ~~two~~ people on board," said Bacon. "We lost radar ~~contact~~ with the helicopter about 9:15 EST (0215 GMT)."

Fig 8

- (a) An OH-58 helicopter, carrying a crew of two, was on a routine training orientation when ~~contact~~ was lost at about 11:30 a.m. Saturday (9:30 p.m. EST Friday).
- (b) "There were two people on board," said Bacon. "We ~~lost~~ radar ~~contact~~ with the helicopter about 9:15 EST (0215 GMT)."

Fig 9

- (a) ~~An OH-58 helicopter~~ carrying a crew of two, was on a routine training orientation when contact was ~~lost~~ at about 11:30 a.m. Saturday (9:30 p.m. EST Friday).
- (b) "There were two people on board," said Bacon. "We ~~lost~~ radar contact with ~~the helicopter~~ about 9:15 EST (0215 GMT)."

BAKER BOTTS LLP

DECLARATION FOR UTILITY OR DESIGN PATENT APPLICATION (37 CFR 1.63) <input type="checkbox"/> Declaration Submitted with Initial Filing OR <input checked="" type="checkbox"/> Declaration Submitted after Initial Filing (surcharge (37 CFR 1.16 (e)) required)	Attorney Docket Number	A32550 PCT-USA
	First Named Inventor	Judith L. Klavans
	COMPLETE IF KNOWN	
	Application Number	10/018,108
	Filing Date	Dec. 13, 2001
	Group Art Unit	(Not Yet Assigned)
	Examiner Name	(Not Yet Assigned)

As a below named inventor, I hereby declare that:

My residence, mailing address, and citizenship are as stated below next to my name.

I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled:

SYSTEM AND METHOD FOR DETECTING TEXT SIMILARITY OVER SHORT PASSAGES

(Title of the Invention)

the specification of which

☐ is attached hereto

OR

☒ was filed on (MM/DD/YYYY) 12/13/2001

as United States Application Number or PCT International

Application Number 10/018,108 and was amended on (MM/DD/YYYY) (if applicable).

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims, as amended by any amendment specifically referred to above

I acknowledge the duty to disclose information which is material to patentability as defined in 37 CFR 1.56, including for continuation-in-part applications, material information which became available between the filing date of the prior application and the national or PCT international filing date of the continuation-in-part application.

I hereby claim foreign priority benefits under 35 U.S.C. 119(a)-(d) or (f), or 365(b) of any foreign application(s) for patent, inventor's or plant breeder's rights certificate(s), or 365(a) of any PCT international application which designated at least one country other than the United States of America, listed below and have also identified below, by checking the box, any foreign application for patent, inventor's or plant breeder's rights certificate(s), or any PCT international application having a filing date before that of the application on which priority is claimed.

Prior Foreign Application Number(s)	Country	Foreign Filing Date (MM/DD/YYYY)	Priority Not Claimed	Certified Copy Attached?	
				YES	NO
			<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
			<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
			<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
			<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

☐ Additional foreign application numbers are listed on a supplemental priority data sheet PTO/SB/02B attached hereto:

DECLARATION — Utility or Design Patent Application**Claim for Benefit of Prior U.S. Provisional Application(s)**

I hereby claim the benefit under Title 35, United States Code, § 119(e) of any United States provisional application(s) listed below:

Provisional Application Number	Filing Date

Claim for Benefit of Earlier U.S./PCT Application(s) under 35 U.S.C. 120

(complete this part only if this is a divisional, continuation or C-I-P application)

I hereby claim the benefit under Title 35, United States Code, § 120 of any United States application(s) or PCT international application(s) designating the United States of America that is/are listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior application(s) in the manner provided by the first paragraph of Title 35, United States Code § 112, I acknowledge the duty to disclose information as defined in Title 37, Code of Federal Regulations, Section 1.56 which occurred between the filing date of the prior applications(s) and the national or PCT international filing date of this application:

Application Number	Filing Date	Status (patented, pending, abandoned)
PCT/US00/40238	Dec. 13, 2001	Pending

DECLARATION — Utility or Design Patent Application

Direct all correspondence to. ☒ Customer Number or Bar Code Label 21003 OR ☒ Correspondence address below

Name

Address

City

State

ZIP

Country

Telephone

Fax

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under 18 U.S.C. 1001 and that such willful false statements may jeopardize the validity of the application or any patent issued thereon

NAME OF SOLE OR FIRST INVENTOR: ☐ A petition has been filed for this unsigned inventor1-00
Given Name
(first and middle [if any])
Judith L.Klavans
Family Name
or SurnameInventor's
Signature
Judith L. Klavans

Date 7/30/02

Hastings-on-Hudson
Residence: City XIXNew York
StateU.S.A.
CountryU.S.A.
CitizenshipMailing Address
40 South DriveHastings-on-Hudson
CityNew York
State10706
ZIPU.S.A.
CountryNAME OF SECOND INVENTOR: ☐ A petition has been filed for this unsigned inventor2-00
Given Name
(first and middle [if any])
EleazarEskin
Family Name
or SurnameInventor's
Signature
Eleazar Eskin

Date 7/24/02

Santa Monica
Residence: City CACA
StateU.S.A.
CountryU.S.A.
CitizenshipMailing Address
935 Stanford StreetSanta Monica
CityCA
State90403
ZIPU.S.A.
Country☒ Additional inventors are being named on the 1 supplemental Additional Inventor(s) sheet(s) PTO/SB/02A attached hereto

BAKER BOTTS LLP

Please type a plus sign (+) inside this box → **+**

DECLARATION

ADDITIONAL INVENTOR(S)
Supplemental Sheet
Page 4 of 4

Name of Additional Joint Inventor, if any:

☐ A petition has been filed for this unsigned inventor

Given Name (first and middle [if any])

Family Name or Surname

Vasileios

Hatzivassiloglou

Inventor's
Signature

V Hatzivassiloglou

Date

7/18/02

New York

Residence: City

NY

State

U.S.A.

Country

Greece

Citizenship

452 Riverside Drive, Apt. 41

Mailing Address

Mailing Address

New York

City

NY
State

10027

ZIP

U.S.A.

Country

Name of Additional Joint Inventor, if any:

☐ A petition has been filed for this unsigned inventor

Given Name (first and middle [if any])

Family Name or Surname

Inventor's
Signature

Date

Residence: City

State

Country

Citizenship

Mailing Address

Mailing Address

City

State

ZIP

Country

Name of Additional Joint Inventor, if any:

☐ A petition has been filed for this unsigned inventor

Given Name (first and middle [if any])

Family Name or Surname

Inventor's
Signature

Date

Residence: City

State

Country

Citizenship

Mailing Address

Mailing Address

City

State

ZIP

Country